Stochastic shortest path problems with associative criteria(The Development of Information and Decision Processes)

AUTHOR(S):

Ohtsubo, Yoshio

CITATION:

Ohtsubo, Yoshio. Stochastic shortest path problems with associative criteria(The Development of Information and Decision Processes). 数理解析研究所講究録 2006, 1504: 95-104

ISSUE DATE:

2006-07

URL:

http://hdl.handle.net/2433/58496

数理解析研究所講究録
1504巻 2006年 95-104

Stochastic shortest path problems with associative criteria (結合型評価に対する確率的最短路問題)

高知大学理学部数理情報科学科
大坪　義夫（Yoshio Ohtsubo）

Department of Mathematics, Faculty of Science, Kochi University

Abstract

We consider a stochastic shortest path problem with associative criteria in which for each node of a graph we choose a probability distribution over the set of successor nodes so as to reach a given target node optimally. We formulate such a problem as an associative Markov decision processes. We show that an optimal value function is a unique solution to an optimality equation and find an optimal stationary policy. Also we give a value iteration method and a policy improvement method.

Keywords : shortest path problem, Markov decision process, associative criterion, invariant imbedding method, optimality equation, existence of optimal policy.

1. Introduction

For a directed graph with nodes $1, 2, \ldots, K$ and with a cost (length or time) assigned to each arc, a stochastic shortest path problem is to select a probability distribution over all possible successor nodes at each node $i \neq K$ so as to reach a target node K with minimal associative accumulate cost.

Such a stochastic shortest path problem is analyzed by using the general theory of Markov decision processes in many references. Eaton and Zadeh[3] formulated such a problem as a pursuit problem and they showed that the optimal expected total cost is a unique solution to an optimality equation if at least one proper policy exists, and they gave an optimal value by a value iteration method. Derman in [4, 5] considered the problem, where a target state (node) is absorbing, and proved that the problem has an optimal stationary policy and he gave several methods for obtaining optimal solutions. In [18], Sancho formulated Markov decision processes to analyze the problem and gave a policy iteration method. Bertsekas and Tsitsiklis[2] investigated the problem without the cost nonnegativity assumption and proved a natural generalization of the standard result for the deterministic shortest path problem within the framework of undiscounted finite state Markovian decision processes. In all of these, a criterion function is the expected total cost, which we call an additive case.

Also, Ohtsubo[14] considered a minimizing risk models in stochastic shortest path problems as undiscounted finite Markov processes and showed that an optimal value function is a unique sclution to an optimality equation and found an optimal stationary policy by using an invariant imbedding method. General minimizing risk models in discounted Markov decision processes were investigated in White [20], Wu and Lin [21], Ohtsubo and Toyonaga [13, 15] and Ohtsubo [16].

On the other hand, Maruyama in [9 , 10, 11, 12] investigated deterministic shortest path

problems with associative criteria and show the existence and uniqueness of the optimal value. Especially in [11] he obtained a parameterized recursive equation for the class of the problem by using an invariant imbedding technique.

Furthermore the optimization problem for minimum criteria, which is associative, was first introduced by Bellman and Zadeh[1] as decision-making in fuzzy environment, and Iwamoto et al.[6, 7, 8] and Ohtsubo[17] formulateed their optimization problem as finite horizon Markov decision processes and give a optimal policy by using an invariant imbedding approach.

In this paper we concern ourselves with a stochastic shortest path problem with an associative criterion, which is an expected accumulate cost $E_i^\pi[\bigcirc_{n=1}^{\tau} Y_n] = E_i^\pi[Y_0 \circ Y_1 \circ Y_2 \circ \cdots Y_\tau]$ where Y_n is a cost at nth step, $\circ$ is an operator with an associative property satisfying some conditions, τ is a hitting time to the target node K and E_i^π is an expectation operator when the starting node is i and a policy π is used. In Section 2, we give notations and formulate our model as undiscounted finite Markov decision processes with infinite horizon. In Section 3, we prove that the optimal value function is a unique solution to an optimality equation by using an invariant imbedding approach and that it is given by a value iteration method. We also show that there exists an optimal left continuous stationary policy. In Section 4, we give a policy improvement method for obtained a optimal policy.

2. Notations and formulation

In this section we formulate associative models in stochastic shortest path problems as Markov decision Processes $\Gamma = ((X_n), (A_n), (Y_n), p)$ with a discrete time space $N = \{0, 1, 2, \ldots\}$. The state space S is a finite set $\{1, 2, \ldots, K\}$ where K is a target state, and we denote the state at time $n \in N$ by X_n. The action space A is finite and we denote the action at time $n \in N$ by A_n. The cost space E is a finite set $\{y_1, y_2, \ldots, y_\ell\}$, where $E \subset B$ for some subset B of R, and $Y_n \in E$ is a random cost function at time $n \in N$ with $Y_0 = e$, where e is a unit element defined below. We define conditional probability distributions by

$$\begin{aligned} q^a(j|i) &= P(X_{n+1} = j | X_n = i, A_n = a), \\ \hat{q}_{ij}^a(y) &= P(Y_{n+1} = y | X_n = i, X_{n+1} = j, A_n = a) \end{aligned}$$

and set

$$p^a(j, y|i) = q^a(j|i)\hat{q}_{ij}^a(y) = P(X_{n+1} = j, Y_{n+1} = y | X_n = i, A_n = a)$$

for $i, j \in S, a \in A$ and $y \in E$. We use $S_B = S \times B$ as a new state space.

For a binary operator $\circ : R \times R \to R$ and a subset B of R, we assume that

(i) B is closed for the operator $\circ$, that is, $x \circ y \in B$ for any $x, y \in B$,
(ii) the operator $\circ$ is associative, that is, $(x \circ y) \circ z = x \circ (y \circ z)$ $(= x \circ y \circ z$, say) for any $x, y, z \in B$,
(iii) B has a unit element e, that is, $e \in B$ and $x \circ e = e \circ x = x$ for any $x \in B$,
(iv) $(B, \circ)$ is nondecreasing in the sense that $x \leq x \circ y$ and $x \leq y \circ x$ for any $x, y \in B$.

In algebra, $(B, \circ)$ satisfying the condition (i), (ii) and (iii) is called semigroup. On the condition (iv), letting $x = e$, we notice that $y \geq e$ for any $y \in B$. Also we easily see under (i), (ii) and

(iii) that if $x \geq e$ and if $x \circ y \leq x \circ z$ and $y \circ x \leq z \circ x$ when $y \leq z$ for any $x, y, z \in B$ then the condition (iv) holds.

We give several examples in which $(B, \circ)$ satisfies the above conditions (cf. Maruyama[10]).

Example 2.1.

(1) (Additive case). When $x \circ y = x + y$, we have $B = [0, \infty)$ and $e = 0$.

(2) (Multiplicative case). When $x \circ y = Lxy$ for a constant $L > 0$, we have $B = [1/L, \infty)$ and $e = 1/L$.

(3) (Maximum case) When $x \circ y = \max(x, y)$, we have $B = [L, M]$ and $e = L$ for constants $L, M \in R$ such that $-\infty < L < M \leq \infty$.

(4) (Multiplicative-additive case). When $x \circ y = x + y - Lxy$ for a constant $L > 0$, we have $B = [0, 1/L]$ and $e = 0$.

(5) (Fractional case). When $x \circ y = (x + y)/(1 + Lxy)$ for a constant $L > 0$, we have $B = [0, 1/\sqrt{L}]$ and $e = 0$.

Let a stopping time τ be a hitting time to the target state K, that is, τ is the smallest nonnegative integer n such that $X_n = K$ where $\tau = \infty$ if there does not exist such an integer n. Then we define the random reward as a criterion function by

$$Z = \bigcirc_{n=0}^{\tau} Y_n \equiv Y_0 \circ Y_1 \circ \cdots \circ Y_\tau.$$

Then our problem is to minimize the expected reward $E_i^\pi[Z]$ with respect to all policies π.

To simplify the optimization problem, we can redefine the equivalent version of the Markov decision processes as follows. We assume that the target state K is absorbing and cost-free, that is, $q^a(K|K) = 1$ and $\hat{q}^a_{KK}(e) = 1$ and hence $p^a(K, e|K) = 1$ for all $a \in A$. Under this assumption we have

$$Z = \bigcirc_{k=0}^{\infty} Y_k \equiv \lim_{n \to \infty} \bigcirc_{k=0}^{n} Y_k,$$

which exists from the remark of the assumption (iv), where we admit $Z = \infty$.

In order to analysis our problem we also define the random reward for a subproblem by

$$Z_n = \bigcirc_{k=1}^{n} Y_k \equiv Y_0 \circ Y_1 \circ \cdots \circ Y_n, \ n \geq 0,$$

Further we define another random sequence as an imbedded parameter by

$$\Lambda_0 = \lambda, \ \ \Lambda_{n+1} = \Lambda_n \circ Y_{n+1}, \ \ n \geq 0,$$

where λ is a given initial parameter in B.

Let $H_0 = S_B$ and $H_{n+1} = H_n \times A \times S_B$ for each $n \in N$. Then H_n represents the set of all possible histories of the system when the nth action must be chosen, and we denote by θ_n the history at time $n \in N$. A decision rule δ_n for time $n \in N$ is a conditional probability given θ_n: $\delta_n(a_n|h_n) = P(A_n = a_n|\theta_n = h_n)$, where $h_n = (i_0, \lambda_0, a_0, i_1, \lambda_1, \ldots, a_{n-1}, i_n, \lambda_n) \in H_n$ which is a realising value of $\theta_n = (X_0, \Lambda_0, A_0, X_1, \Lambda_1, \ldots, A_{n-1}, X_n, \Lambda_n)$. It is assumed that $\delta_n(A_n \in A|h_n) = 1$ for every history $h_n = (i_0, \lambda_0, a_0, \ldots, i_n, \lambda_n) \in H_n$. We denote by Δ the

set of all decision rules. A policy π is an infinite sequence of decision rules $(\delta_n, n \geq 0) = (\delta_0, \delta_1, \delta_2, \ldots, \delta_n, \ldots)$. We denote by C the set of all such policies.

A policy $\pi = (\delta_n, n \geq 0)$ is said to be Markov when the decision rule δ_n is a function of $(X_n, \Lambda_n) = (i_n, \lambda_n)$ for every $n \in N$. We denote the set of such decision rules by Δ_M and the set of all Markov policies by C_M. Also, a policy π is called a deterministic Markov policy if π is Markov and $\delta_n(a|i, \lambda) = 1$ for some $a \in A$. We write $\delta_n(i, \lambda) = a$ for such a decision rule δ_n and we denote by Δ_D the set of such decision rules. We also denote the set of all deterministic Markov policies by C_D. When $\delta_n = \delta \in \Delta_D$ for all $n \in N$, we write $\pi = \delta^\infty$, which is called a stationary policy, and we denote the set of all stationary policies by C_D^s.

We denote by $E_i^\pi[Z]$ the conditional expectation of Z given an initial state $X_0 = i$ and a policy $\pi \in C$. Since the random variable Z depends upon not only i and π but also λ, we may sometimes use a conditional probability $P_{(i,\lambda)}^\pi(\cdot)$ and an expectation $E_{(i,\lambda)}^\pi(\cdot)$. Through this paper we assume that $P_{(i,\lambda)}^\pi(X_n = K \text{ for some } n \geq 0) = P_{(i,\lambda)}^\pi(\tau < \infty) = 1$ for every stationary policy $\pi \in C_D^s$ and each $(i, \lambda) \in S_B$, that is, the states $1, 2, \ldots, K-1$ are transient when we use any policy $\pi \in C_D^s$. Thus we easily see that $P_{(i,\lambda)}^\pi(Z < \infty) = 1$ for all $\pi \in C_D^s$ and each $(i, \lambda) \in S_B$. This is analogous to a condition given in Ohtsubo[16].

A decision rule $\delta \in \Delta_D$ is said to be left continuous (on B) if for each $(i, \lambda) \in S_B$ there is a positive real number μ such that $\delta(i, \lambda) = \delta(i, \lambda - u)$ for all $u : 0 \leq u < \mu$ such that $\lambda - u \in B$. A policy $\pi = \delta^\infty \in C_D^s$ is said to be left continuous if the decision rule δ is left continuous.

In order to analysis our model, we denote criterion functions for finite and infinite horizon cases by

$$F_n^\pi(i, \lambda) = E_i^\pi[\lambda \circ Z_n], \quad F^\pi(i, \lambda) = E_i^\pi[\lambda \circ Z],$$

respectively, for each $(i, \lambda) \in S_B$ and $\pi \in C$. When $n = 3$, the explicit form of the expectation $F_3^\pi(i_1, \lambda)$ is

$$\begin{aligned} E_{i_1}^\pi[\lambda \circ Z_3] \;=\; & \sum_{a_1 \in A} \sum_{y_1 \in E} \sum_{i_2 \in S} \sum_{a_2 \in A} \sum_{y_2 \in E} \sum_{i_3 \in S} \sum_{a_3 \in A} \sum_{y_3 \in E} \sum_{i_4 \in S} (\lambda \circ y_1 \circ y_2 \circ y_3) \\ & \times p^{a_3}(i_4, y_3|i_3)\delta_2(a_3|i_1, \lambda, a_1, i_2, \lambda \circ y_1, a_2, i_3, \lambda \circ y_1 \circ y_2) \\ & \times p^{a_2}(i_3, y_2|i_2)\delta_1(a_2|i_1, \lambda, a_1, i_2, \lambda \circ y_1) \\ & \times p^{a_1}(i_2, y_1|i_1)\delta_0(a_1|i_1, \lambda) \end{aligned}$$

for $(i_1, \lambda) \in S_B$ and $\pi = (\delta_0, \delta_1, \delta_2, \cdots) \in C$. We also define optimal value functions F_n^* and F^* for finite and infinite horizon cases by, respectively,

$$F_n^*(i, \lambda) = \inf_{\pi \in C} F_n^\pi(i, \lambda), \quad F^*(i, \lambda) = \inf_{\pi \in C} F^\pi(i, \lambda).$$

Then we notice that optimal value in the original problem is

$$F^*(i, e) = \sup_{\pi \in C} F^\pi(i, e) = \sup_{\pi \in C} E_i^\pi[Z],$$

since e is the unit element. A policy π is said to be optimal if $F^*(i, \lambda) = F^\pi(i, \lambda)$ for every $(i, \lambda) \in S_B$.

We define the following sets of functions: let $\mathcal{F}$ be the set of functions F from S_B into B such that $F(i, \lambda)$ is measurable on B for each $i \in S$ and $F(i, \lambda) \geq \lambda$, let $\mathcal{F}_B$ be the set of functions $F \in \mathcal{F}$ such that $F(\cdot, \lambda)$ is bounded for each $\lambda \in B$, and let $\mathcal{F}_\ell$ be the set of functions $F \in \mathcal{F}$

such that $F(i,\cdot)$ is nondecreasing and left continuous on B for each $i \in S$. In Theorem 3.1 it is shown that $F^* \in \mathcal{F}_\ell$. However, it is not necessarily true that $F^\pi \in \mathcal{F}_\ell$ for each $\pi \in C$.

We finally define operators T^a, T^δ and T from $\mathcal{F}$ into itself as follows. For $F \in \mathcal{F}$, $(i,\lambda) \in S_B$, $a \in A$ and $\delta \in \Delta_M$,

$$T^aF(i,\lambda) = \sum_{j\in S}\sum_{y\in E} F(j,\lambda\circ y)p^a(j,y|i),$$

$$T^\delta F(i,\lambda) = \sum_{a\in A} T^aF(i,\lambda)\delta(a|i,\lambda),$$

$$TF(i,\lambda) = \inf_{\delta\in\Delta} T^\delta F(i,\lambda) = \min_{a\in A} T^aF(i,\lambda).$$

We also define operators T^n by $T^1 = T$ and $T^{n+1} = T(T^n), n \geq 1$. Similarly, $(T^\delta)^n$ is defined for $\delta \in \Delta_M$. In all argument, for $F, G \in \mathcal{F}$, $F \geq G$ means that $F(i,\lambda) \geq G(i,\lambda)$ for all $(i,\lambda) \in S_B$.

3. Optimal value and optimal policy

In this section we prove that the optimal value function is a unique solution to an optimality equation and we give a value iteration method. These results are an associative extension of Eaton and Zadeh[3], Derman[4, 5], and Bellman and Zadeh[1], and a stochastic one of Maruyama[11]. We also show that there exists an optimal left continuous policy.

We first give fundamental lemmas below.

Lemma 3.1.

(i) *For $F, G \in \mathcal{F}$ and $\delta \in \Delta$, $T^\delta F - T^\delta G = T^\delta(F-G)$.*
(ii) *If $F, G \in \mathcal{F}$ and $F \geq G$, then $T^aF \geq T^aG$ for each $a \in A$, $T^\delta F \geq T^\delta G$ for each $\delta \in \Delta$ and $TF \geq TG$.*
(iii) *If $G \in \mathcal{F}_\ell$, then $T^aG \in \mathcal{F}_\ell$ for any $a \in A$. Also, T is an operator from $\mathcal{F}$ (or $\mathcal{F}_\ell$) into itself.*
(iv) *If $G_n \in \mathcal{F}_\ell$ and $G_n \leq G_{n+1}$ for each $n \geq 0$, then $\lim_{n\to\infty} G_n \in \mathcal{F}_\ell$.*

We easily see that for each $F \in \mathcal{F}$, there is a measurable decision rule $\delta \in \Delta_D$ satisfying $TF = T^\delta F$, since TF is measurable and A is finite.

Furthermore, the following lemma is important for main theorems.

Lemma 3.2. *For each $F \in \mathcal{F}_\ell$, there exists a left continuous decision rule $\delta \in \Delta_D$ satisfying $TF = T^\delta F$.*

For any $\pi = (\delta_n, n \geq 0) \in C$ and a given history $(i,\lambda,a) \in S_B \times A$, the cut-head policy of π to (i,λ,a) is defined by ${}^1\pi^{(i,\lambda,a)} = (\delta_n^{(i,\lambda,a)}, n \geq 0)$ where $\delta_n^{(i,\lambda,a)}(\cdot|h_n) = \delta_{n+1}(\cdot|(i,\lambda,a),h_n)$ for every $h_n \in H_n$ and each $n \geq 0$. Then we see that ${}^1\pi^{(i,\lambda,a)} \in C$ for a fixed (i,λ,a). For the sake of simplicity we use a notation:

$$T^{\delta_0}F^{{}^1\pi}(i,\lambda) = \sum_{a\in A}\delta_0(a|i,\lambda)\sum_{j,,y} F^{{}^1\pi^{(i,\lambda,a)}}(j,\lambda\circ y)p^a(j,y|i)$$

for each $\pi = (\delta_n, n \geq 0) \in C$ and $(i,\lambda) \in S_B$.

Lemma 3.3. *Let $\pi = (\delta_n, n \geq 0) \in C$ be arbitrary. For each $n \geq 0$, $F_{n+1}^\pi = T^{\delta_0}F_n^{{}^1\pi}$ and $F^\pi = T^{\delta_0}F^{{}^1\pi}$. Especially, $F^\pi = T^\delta F^\pi$ when $\pi = \delta^\infty \in C_D^s$.*

We next give fundamental properties for optimal value functions of finite and infinite horizon

cases.

Theorem 3.1. We have the following:

(i) *For each* $n \geq 0$, $F_n^* \in \mathcal{F}_\ell$ *and* $\{F_n^*, n \geq 0\}$ *satisfies equations* :

$$F_0^*(i,\lambda) = \lambda, \ (i,\lambda) \in S_B, \quad F_n^* = TF_{n-1}^*, \quad n \geq 1.$$

(ii) *For each* $n \geq 0$, *there exists a left continuous policy* $\pi \in C_D$ *such that* $F_n^* = F_n^\pi$.
(iii) *For each* $n \geq 0$, $F_n^* \leq F_{n+1}^* \leq \lim_{n\to\infty} F_n^* \leq F^*$ *and* $\lim_{n\to\infty} F_n^* \in \mathcal{F}_\ell$.

Remark. On the statement (iii) we have $\lim_{n\to\infty} F_n^* = F^*$ under some conditions, which we will prove in Theorem 3.2.

From Theorem 3.1, we have $F_n^* = T^n F_0^*$ for each $n \geq 0$. In order to prove that $F^* = \lim_{n\to\infty} F_n^*$, we need the following important lemma.

Lemma 3.4. *Let* $\pi = \delta^\infty \in C_D^s$ *be a policy satisfying the condition that for each* $(i,\lambda) \in S_B$ *there is a constant* $M > 0$ *such that* $P_{(i,\lambda)}^\pi(\lambda \circ Z \leq M) = 1$.

(i) *Let* $F, G \in \mathcal{F}_B$. *If* $F - G \leq T^\delta(F - G)$ *on* $\{K\}^c \times B$ *and* $F = G$ *on* $\{K\} \times B$, *then* $F \leq G$ *on* S_B.
(ii) F^π *is the unique solution in* $\mathcal{F}_B$ *to equation* $F = T^\delta F$ *with* $F(K,\lambda) = \lambda$ *for every* $\lambda \in B$.

Now we are in a position to give a main theorem.

Theorem 3.2. *Suppose that there exists at least one policy* $\sigma \in C$ *such that for each* $(i,\lambda) \in S_B$ *there is a constant* $M > 0$ *such that* $P_{(i,\lambda)}^\sigma(\lambda \circ Z \leq M) = 1$.

(i) $F^* = \lim_{n\to\infty} F_n^*$.
(ii) F^* *is the unique solution in* $\mathcal{F}_B$ *to* $F = TF$ *with* $F(K,\lambda) = \lambda$ *for every* $\lambda \in B$.
(iii) *There exists a left continuous policy* $\pi = \delta^\infty \in C_D^s$ *satisfying* $F^* = T^\delta F^*$ *on* $\{K\}^c \times B$ *and* π *is optimal.*

In order to consider special cases, we define new policy spaces below. Let a policy space Π be the set of all policies $\pi = (\delta_n, n \geq 0) \in C$ such that δ_n is not depend upon parameters $\lambda_0, \lambda_1, \ldots, \lambda_k, \ldots$ for every $n \geq 0$. Similarly, we defien Π_M, Π_D and Π_D^s corresponding to C_M, C_D and C_D^s, respectively. For example, $\pi = \delta^\infty \in \Pi_D^s$ is a policy such that $\delta(i) = a$ for each $i \in S$ and some $a \in A$. These are usual policy spaces defined on Markov decision processes with additive criteria (cf. White[19]).

Corollary 3.1. *Suppose that there exists at least one policy* $\sigma \in \Pi$ *such that for each* $i \in S$ *there is a consiant* $M > 0$ *such that* $P_i^\sigma(Z \leq M) = 1$.

(i) *In an additive case, that is,* $x \circ y = x + y$, *there is an optimal policy* $\pi = \delta^\infty \in \Pi_D^s$ *such that* $F^*(i,0) = T^\delta F^*(i,0)$ *for each* $i \in \{K\}^c$ *and* $F^*(K,0) = 0$.
(ii) *In an multiplicative case, that is,* $x \circ y = Lxy$ *for a constant* $L > 0$, *there is an optimal policy* $\pi = \delta^\infty \in \Pi_D^s$ *such that* $F^*(i,1/L) = T^\delta F^*(i,1/L)$ *for each* $i \in \{K\}^c$ *and* $F^*(K,1/L) = 1/L$.

(iii) *In an multiplicative-additive case, that is,* $x \circ y = x + y - Lxy$ *for a constant* $L > 0$, *there is an optimal policy* $\pi = \delta^\infty \in \Pi_D^s$ *such that* $F^*(i,0) = T^\delta F^*(i,0)$ *for each* $i \in \{K\}^c$ *and* $F^*(K,0) = 0$.

From Theorems 3.1 and 3.2 we see that a value iteration is given by $F^* = \lim_{n\to\infty} T^n F_0^*$ where $F_0^*(i,\lambda) = \lambda$ for each $(i,\lambda) \in S_B$. We give another value iteration in the following theorem.

Theorem 3.3. *Suppose that there is a policy* $\sigma \in C$ *such that for each* $(i,\lambda) \in S_B$ *there is a constant* $M > 0$ *such that* $P^\sigma_{(i,\lambda)}(\lambda \circ Z \le M) = 1$.

Let $G \in \mathcal{F}$ *be a function satisfying* $G \le F^*$. *Then* $\{T^n G\}$ *converges and* $\lim_{n\to\infty} T^n G = F^*$.

4. Policy iteration method

In this section we consider a policy space iteration procedure in our model as follows:

(i) Select an initial policy $\pi_0 = (\delta_0)^\infty \in C_D^s$.
(ii) At step n, assume that we have a policy $\pi_n = (\delta_n)^\infty \in C_D^s$ and solve the equation $F = T^{\delta_n} F$ with $F(K,\lambda) = \lambda$ for every $\lambda \in B$ to give a function $F^{\pi_n} \in \mathcal{F}$.
(iii) If $T^{\delta_n} F^{\pi_n} = TF^{\pi_n}$, stop the procedure. If $T^{\delta_n} F^{\pi_n} \neq TF^{\pi_n}$, go the next step.
(iv) Find a new policy $\pi_{n+1} = (\delta_{n+1})^\infty \in C_D^s$ by $T^{\delta_{n+1}} F^{\pi_n} = TF^{\pi_n}$.
(v) Return to step (ii) replacing n by $n+1$.

From Lemma 3.4(ii) we can uniquely solve the equations in $\mathcal{F}$ at step (ii) under some conditions. We have the following convergence theorem.

Theorem 4.1. *Suppose that there exists at least one policy* $\sigma \in C$ *such that for each* $(i,\lambda) \in S_B$ *there is a constant* $M > 0$ *such that* $P^\sigma_{(i,\lambda)}(\lambda \circ Z \le M) = 1$.

(i) The sequence $\{F^{\pi_n}\}$ is nonincreasing and converges to F^*.
(ii) If $T^{\delta_n} F^{\pi_n} = TF^{\pi_n}$, then F^{π_n} is the optimal value and $\pi_n = (\delta_n)^\infty \in C_D^s$ is an optimal policy.

5. Examples

We first consider an example of a maximum case and get optimal value and optimal policy by the policy iteration method.

Example 5.1. Let $x \circ y = \max(x,y)$. Let $S = \{1,2,3\}$ be a state space and 3 be a target node. Assume that the state 3 is absorbing and cost-free. Also let $A = \{a_1, a_2\}$ be an action space. We give the probability distributions by

$$p^{a_1}(2,2|1) = \frac{2}{3}, \quad p^{a_1}(3,2|1) = \frac{1}{3},$$
$$p^{a_1}(3,6|2) = p^{a_2}(2,4|1) = 1,$$
$$p^{a_2}(2,8|2) = p^{a_2}(3,3|2) = \frac{1}{2}.$$

Then we have $B = [2,8]$ and $e = 2$. We consider a policy space procedure to give an optimal policy. Let $\pi_0 = (\delta_0)^\infty \in C_D^s$ be an initial policy such that $\delta_0(i,\lambda) = a_1$ for every $(i,\lambda) \in S_B$.

Solving the equation $F = T^{\delta_0}F$ with $F(3,\lambda) = \lambda$ for every $\lambda \in B$, we have

$$F^{\pi_0}(2,\lambda) = \begin{cases} 6 & (2 \le \lambda \le 6) \\ \lambda & (6 < \lambda \le 8) \end{cases}, \quad F^{\pi_0}(1,\lambda) = \begin{cases} \frac{1}{3}\lambda + 4 & (2 \le \lambda \le 6) \\ \lambda & (6 <\le \lambda \le 8) \end{cases}$$

We now see that $T^{\delta_0}F^{\pi_0} \neq TF^{\pi_0} = \min(T^{a_1}F^{\pi_0}, T^{a_2}F^{\pi_0})$, since

$$T^{a_1}F^{\pi_0}(2,\lambda) = F^{\pi_0}(2,\lambda), \quad T^{a_2}F^{\pi_0}(2,\lambda) = \begin{cases} \frac{11}{2} & (2 \le \lambda \le 3) \\ \frac{\lambda}{2} + 4 & (3 < \lambda \le 8) \end{cases}.$$

Next, using $T^{\delta_1}F^{\pi_0} = TF^{\pi_0}$, we give a policy $\pi_1 = (\delta_1)^\infty \in C_D^s$ by

$$\begin{aligned} &\delta_1(3,\lambda) = a_1 \\ &\delta_1(2,\lambda) = \begin{cases} a_2 & (2 \le \lambda \le 4) \\ a_1 & (4 < \lambda \le 8) \end{cases}, \\ &\delta_1(1,\lambda) = a_1. \end{aligned}$$

By solving $F = T^{\delta_1}F$ with $F(3,\lambda) = \lambda$, F^{π_1} is given by

$$F^{\pi_1}(2,\lambda) = \begin{cases} \frac{11}{2} & (2 \le \lambda \le 3) \\ \frac{\lambda}{2} + 4 & (3 < \lambda \le 4) \\ 6 & (4 < \lambda \le 6) \\ \lambda & (6 < \lambda \le 8) \end{cases}, \quad F^{\pi_1}(1,\lambda) = \begin{cases} \frac{1}{3}\lambda + \frac{11}{3} & (2 \le \lambda \le 3) \\ \frac{2}{3}\lambda + \frac{8}{3} & (3 < \lambda \le 4) \\ \frac{1}{3}\lambda + 4 & (4 < \lambda \le 6) \\ \lambda & (6 < \lambda \le 8) \end{cases}$$

We can easily check that $T^{\delta_1}F^{\pi_1}(i,\lambda) = TF^{\pi_1}(i,\lambda)$ for every $(i,\lambda) \in S_B$. Thus we stop the procedure. From Theorem 4.1 we obtain the optimal value $F^* = F^{\pi_1}$ and an optimal policy $\pi_1 = (\delta_1)^\infty$. Therefore, since $e = 2$, we have optimal value in the original problem as follows:

$$F^*(1,2) = \frac{13}{3}, \quad F^*(2,2) = \frac{11}{2}, \quad F^*(3,2) = 2.$$

We next consider an example of a multiplicative case.

Example 5.2. Let $x \circ y = xy$. Let $S = \{1,2,3\}$ be a state space and 3 be a target node. Assume that the state 3 is absorbing and cost-free. Also let $A = \{a_1, a_2\}$ be an action space. We give the probability distributions by

$$\begin{aligned} &p^{a_1}(2,2|1) = \frac{2}{3}, \quad p^{a_1}(3,2|1) = \frac{1}{3}, \\ &p^{a_1}(3,6|2) = p^{a_2}(2,4|1) = 1, \\ &p^{a_2}(2,5|2) = \frac{1}{16}, \quad p^{a_2}(3,3|2) = \frac{15}{16}. \end{aligned}$$

Then we have $B = [1,\infty)$ and $e = 1$. From Corollary 3.1, We may put $\lambda = e = 1$ to analysis the multiplicative case, but we use λ. We consider a policy space procedure to give an optimal value and an optimal policy. Let $\pi_0 = (\delta_0)^\infty \in C_D^s$ be an initial policy such that $\delta_0(i,\lambda) = a_1$ for every $(i,\lambda) \in S_B$. Solving the equation $F = T^{\delta_0}F$ with $F(3,\lambda) = \lambda$ for every $\lambda \in B$, we have

$$F^{\pi_0}(2,\lambda) = 6\lambda, \quad F^{\pi_0}(1,\lambda) = \frac{26}{3}\lambda$$

We now see that $T^{\delta_0}F^{\pi_0} \neq TF^{\pi_0}$, since

$$T^{a_1}F^{\pi_0}(2,\lambda) = F^{\pi_0}(2,\lambda) = 6\lambda, \quad T^{a_2}F^{\pi_0}(2,\lambda) = \frac{75}{16}\lambda$$

Next, using $T^{\delta_1}F^{\pi_0} = TF^{\pi_0}$, we give a policy $\pi_1 = (\delta_1)^\infty \in C_D^s$ by

$$\delta_1(3,\lambda) = a_1, \quad \delta_1(2,\lambda) = a_2, \quad \delta_1(1,\lambda) = a_1.$$

By solving $F = T^{\delta_1}F$ with $F(3,\lambda) = \lambda$, F^{π_1} is given by

$$F^{\pi_1}(2,\lambda) = \frac{45}{11}\lambda, \quad F^{\pi_1}(1,\lambda) = \frac{112}{33}\lambda.$$

We can easily check that $T^{\delta_1}F^{\pi_1}(i,\lambda) = TF^{\pi_1}(i,\lambda)$ for every $(i,\lambda) \in S_B$. Thus we stop the procedure. We obtain the optimal value $F^* = F^{\pi_1}$ and an optimal policy $\pi_1 = (\delta_1)^\infty$. Therefore, since $e = 1$, we have optimal value in the original problem as follows:

$$F^*(1,1) = \frac{112}{33}, \quad F^*(2,1) = \frac{45}{11}, \quad F^*(3,1) = 1.$$

Reference

[1] Bellman, R.E. and Zadeh, L.A. (1970) Decision-making in a fuzzy environment, *Management Science*, 17, B141-B164

[2] Bertsekas, D.P. and Tsitsiklis, J.N. (1991) An analysis of stochastic shortest path problems. *Math. Oper. Res.*, 16, 580-595

[3] Eaton, J.H. and Zadeh, L.A. (1962) Optimal pursuit strategies in discrete-state probabilistic systems. *Trans. ASME Ser. D, J. Basic Eng.*, 84, 23-29

[4] Derman, C. (1962) On sequential decisions and Markov chains. *Manage. Sci.*, 9, 16-24

[5] Derman, C. (1970) Finite state Markovian decision processes. Academic Press, New York

[6] Iwamoto, S. and Fujita, T. (1995) Stochastic decision-making in a fuzzy environment, *J. Operations Research Society of Japan*, 38, 467-482

[7] Iwamoto, S., Tsurusaki, K. and Fujita, T. (1999) Conditional decision-making in fuzzy environment, *J. Operations Research Society of Japan*, 42, 198-218

[8] Iwamoto, S., Tsurusaki, K. and Fujita, T. (2001) On Markov policies for minimax decision processes, *J. Math. Anal. Appl*, 253, 58-78

[9] Maruyama, Y. (1997) On associative shortest path problems, *Bulletin of Informatics and Cybernetics*, 29, 67-81

[10] Maruyama, Y. (1999) Associative shortest and longest path problems, *Bulletin of Informatics and Cybernetics*, 31, 147-163

[11] Maruyama, Y. (1999) An invariant imbedding approach to associative shortest path problems, *Math. Japonica*, 50, 469-480

[12] Maruyama, Y. (1999) Duality theorems in parametric associative optimal path problems, *Asia-Pacific J. Operations Research*, 17, 149-168

[13] Ohtsubo, Y. and Toyonaga, K. (2002) Optimal policy for minimizing risk models in Markov decision processes, *J. Math. Anal. Appl*, 271, 66-81

[14] Ohtsubo, Y. (2003) Minimizing risk models in stochastic shortest path problems, *Mathematical Methods of Operations Research*, 57, 79-88

[15] Ohtsubo, Y. and Toyonaga, K. (2004) Equivalence classes for optimizing risk models in Markov decision processes, *Mathematical Methods of Operations Research*, 60, 239-250

[16] Ohtsubo, Y. (2004) Optimal threshold probability in undiscounted Markov decision processes with a target set, *Applied Mathematics and Computation*, 149, 519-532

[17] Ohtsubo, Y. (2006) Multistage Markov decision processes with minimum criteria of random rewards, *Bulletin of Informatics and Cybernetics*, to appear

[18] Sancho, N.G.F. (1985) Routing problems and Markovian decision processes. *J. Math. Anal. Appl.*, 105, 76-83

[19] White, D.J. (1993) Markov decision processes, John Wiley, New York

[20] White, D.J. (1993) Minimizing a threshold probability in discounted Markov decision processes, *J. Math. Anal. Appl*, 173, 634-646

[21] Wu, C. and Lin, Y. (1999) Minimizing risk models in Markov decision processes with policies depending on target values, *J. Math. Anal. Appl*, 231, 47-67